INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2013 EDITION

SYSTEM DRIVERS ABSTRACT

THE ITRS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

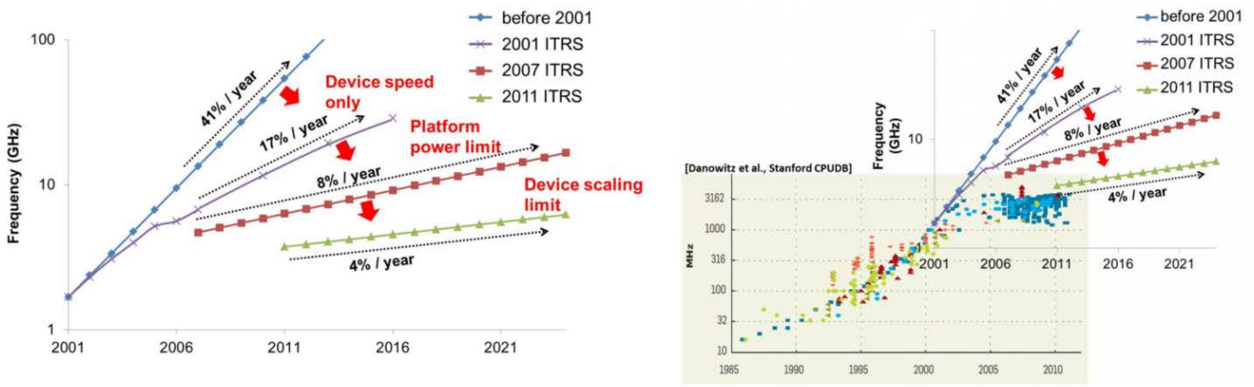
Table of Contents

List of Figures

System Drivers Summary

1. Design Capability Gap

A crucial observation from product data in recent years is that transistor density in actual products has not scaled as would have been expected according to Moore’s Law. Figure SYSD1(a) shows that even as lithography has delivered “available” Moore’s Law scaling (i.e., geometric scaling) per the ITRS roadmap at least through the year 2013, “realized” transistor density scaling has since 2007 slowed to 1.6× per node instead of the traditional 2× per node. The gap between “available” density scaling from patterning pitch, versus “realizable” scaling in actual products, is a clear challenge to the validity of Moore’s Law. Explaining and deconstructing this design capability gap – as a consequence of reliability constraints, variability in process and operating conditions, signoff analysis pessimism, foundry models, design architectures, and, yes, lithography – is critical to future resumption of Moore’s-Law scaling of value.



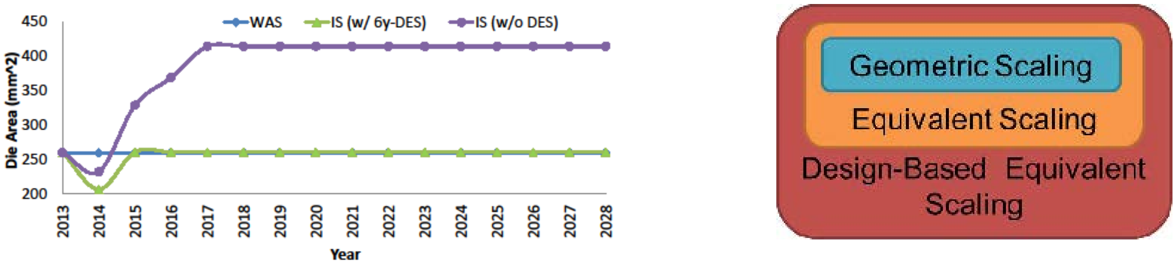
(a) (b)

Figure SYSD1 (a) Evolution of the ITRS frequency roadmap. (b) Overlay of the ITRS frequency roadmap with data from the Stanford CPUDB repository.

To compound the design capability gap, even geometric scaling of Mx pitch is now slowing. Daunting challenges to pitch scaling have arisen from resistivity and manufacturability of damascene copper interconnects, poor design-level ROI from new technologies with heavily restricted layout ground rules, and wide (pessimistic) parasitic extraction corners in multi-patterning technologies. Survey data, physical analysis of recent MPU and SOC products, and announced foundry offerings all point to a slowdown of Mx pitch scaling to a three-year cycle in the next two technology nodes, as opposed to the two-year cycle projected in the 2011 ITRS roadmap. This slowdown, depicted in Figure SYSD1(b), is potentially not restricted to logic products alone. For example, the roadmap for contacted poly half-pitch in NAND flash products will likely have both a “2D” version (18nm in 2013, scaling to 12nm in 2022) and a “3D” version (64nm in 2013, scaling to 26nm in 2022) – either of which allows the product capacity to double every two years. The latter trajectory, like the slowdown of Mx pitch scaling, relaxes the requirements for patterning technologies, and potentially implies a lessening criticality of lithography (or, acknowledgment of risks and costs of EUV, quad-patterning, etc.) in the semiconductor roadmap.

2. Design Equivalent Scaling

We observe that the slowing of Mx pitch scaling introduces a further gap in Moore’s-Law scaling. Specifically, if transistor counts continue to scale at 1.6× per node in order to deliver improved product value, then a slowdown of pitch scaling leads to an explosion of die area as shown by the purple line in Figure SYSD2(a). Near-term compensation of the slowdown in density may be afforded by *design-based equivalent scaling* (DES) (Figure SYSD2(b)) which, like equivalent scaling, achieves non-geometric enhancements of performance, density, and other key value metrics. Examples of DES span error-correcting codes to improve memory reliability, double patterning-aware design techniques that reduce design guardband, clock gating, adaptive voltage and frequency scaling to reduce design margin, etc.



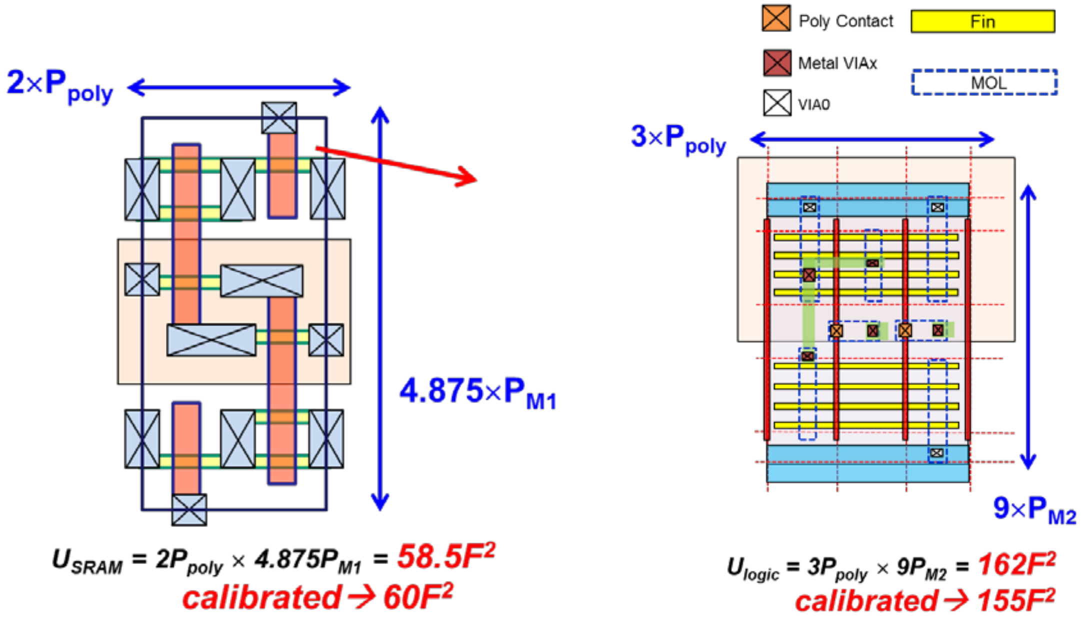
(a)

(b)

Figure SYSD2 (a) Area explosion of MPU-HP (b) Illustration of different scaling approaches

The green line in Figure SYSD2(a) shows the potential impact of DES in a regime of slowed geometric scaling, as transistor counts continue to increase by 1.6× per node. It may be (optimistically) projected that for server and desktop processors (MPU), DES can recover one entire node of Moore's-Law scaling from 2013 to 2019; for processors in SOCs, DES can recover one node of scaling from 2013 to 2020. Put another way, DES can potentially scale down the area of logic overhead (wasted space in logic) to 0.63× its present levels over the next six years, so as to meet the 1.6× transistor density growth requirement and rescue Moore's Law over this near-term time frame.

3. Updated A-Factors



(a)

(b)

Figure SYSD3 (a) New canonical layout for SRAM (FinFET) (b) New canonical layout for NAND2 cell (FinFET)

Through the end of the roadmap, a 2-input FinFET NAND gate is assumed to lay out in 9 × 3 grids (Figure SYSD3(b)), where the vertical dimension is in units of contacted local metal (Metal 2) pitch (PM2 = 2.0 × F), and the horizontal dimension is in unit of contacted poly pitch (P_{poly} = 3 × F). A FinFET 6-transistor SRAM bit cell is assumed to lay out in

2×5 grids (Figure SYSD3(a)), where the vertical dimension is in units of contacted poly pitch (P_{poly} = 3 × F), and the horizontal dimension is in units of contacted Metal1 pitch (PM1=2 × F). In other words, the average gate occupies (9 × 2.0F) × (3 × 3F) = $162F^2$ and the average SRAM bitcell occupies (2 × 3F) × (5 × 2F) = $60F^2$. After fitting with data from production libraries, 155 is chosen for the logic A-factor. The A-factor of 155 (logic) is reduced from the value of 175 used in previous ITRS editions. This is due to the advance of the Mx density and the introduction of middle-of-line (MOL) technology.

4. MPU MODEL SUMMARY OF CHANGES

A major change to the MPU System Driver is that the MPU-PCC will be removed from the roadmap; it is being replaced by SOC-CP. Devices with connections, such as netbooks or tablets, are equipped with mobile SOCs instead of traditional MPUs. Therefore, the MPU-PCC product class is no longer a significant driver in this segment. Based on the introduction of the design capability gap as noted above, the MPU model has been updated accordingly. The details of changes are mentioned in the list of bullets below. To summarize: MPU area will remain the same as in the 2013 roadmap, but the overhead scaling will no longer be constant, so as to compensate discrepancies between the scaling rates of die contents and device/interconnect geometries. Since there will be a stall of M1HP scaling in the next technology node, design-based equivalent scaling (DES) is integrated into the new MPU model (roadmap) to mitigate this scaling crisis. The basic approach to estimate MPU power is similar to previous modeling, but with different design parameters stemming from the design capability gap and DES.

- **[NEW]** Node: M1 half-pitch scales 0.5× / 3 years from 2013 through 2028
- *Die area: **Constant** (= $260mm^2$ (MPU-HP) and $140mm^2$ (MPU-CP) in 2013)
- Transistor (Tx) density scaling: **1.6× / node (node = 2years** till 2019**, 3years** from 2019 onwards)
- **[NEW]** A-factor
    - Logic (NAND2): **155**
    - SRAM: **60** (bulk and FinFET)
- Tx scaling
    - #Tx (Logic + SRAM): **1.6× / node** (= 7.14B (MPU-HP) and 2.54B (MPU-CP) in 2013)
    - #Logic Tx: **1.6× / node** (= 3.68B (MPU-HP) and 1.32B (MPU-CP) in 2013)
    - #Cores: **1.26× / node** (= 8 (MPU-HP) and 4 (MPU-CP) in 2013)
    - #Logic Tx per core: **1.26× / node** (= 0.46B (MPU-HP) and 0.33B (MPU-CP) in 2013)
    - # SRAM Tx: **1.6× / node** (= 3.46B (MPU-HP) and 1.22B (MPU-CP) in 2013)
- **[NEW]** Overheads
    - Logic overhead ($O_{eq-logic}$)[1]: **1.4 in 2013, scales at 1.26× / 3 years**
    - SRAM overhead (O_{SRAM})[2]: **1.3 from 2013 to 2019, scales at 1.26× / node from 2020**
    - Integration overhead ($O_{integration}$)[3]: **1.235** (fixed throughout roadmap)
- **[NEW]** Design-based equivalent scaling for logic
    - 6y-DES for Logic: 1.00 in 2013, scales at **0.93× / year** from 2013 to 2019 (to recover one node of Tx scaling at the end of 6 years)**, 0.63** from 2020 onwards
- Power

[1] Overhead due to whitespace and PDN for logic, uncore elements (e.g., memory controller, IO interfaces, accelerators) and pitch relaxation (for reliability, high frequency operation, use of at least 6X - 12X cell sizes instead of 1X, complex gates (e.g., FFs, MUXes)).

[2] Overhead due to peripheral logic and whtespace for SRAM bitcell, sizing of PU/PD for read/write stability, write assist, reliability and manufacturability increases SRAM overhead beyond 2020.

[3] Overhead due to whitespace for integration of core, uncore, PLL and analog logic blocks.

  - Frequency: **1.04× / year** (same as ITRS 2011) (= 5.5GHz (MPU-HP) and 2.0GHz (MPU-CP) in 2013)
  - Activity factor: **0.95× / year** (same as ITRS 2011) (=0.10 in 2013 (MPU-HP and MPU-CP))
  - Ratio of low-Vt cells (β): **0.1** (same as ITRS 2011)
  - Derated activity factor for non-critical paths (α'): **0.33** (same as ITRS 2011)

5. SOC MODEL SUMMARY OF CHANGES

The basic methodology to model SOC is similar to the previous ITRS edition. The major change to the SOC System Driver class is that the SOC-CS driver is removed in the 2013 roadmap. This is because the boundary between game consoles and mobile devices has become vague. In 2013, we retain only SOC-CP in the SOC segment due to the strong growth in the mobile device segment. The design capability gap and DES are integrated within the new SOC-CP model, as was also done with the MPU model. Power consumption estimates for SOC-CP are now made using a scenario-based approach, since key functional blocks (GPU, RF, and multimedia IPs) have very extreme switching activity discrepancies in different scenarios.

- Node: M1 half-pitch scales 0.5× / 2 years till 2017; 0.5× / 3years beyond 2017 till 2028
- *Die area: **Constant** (= 140mm^2 in 2013)
- Transistor (Tx) density scaling: **1.6× / node (node = 2years** till 2019**, 3years** from 2019 onwards)
- Tx scaling
  - #Tx (Logic + SRAM): **1.6× / node** (= 2.4B in 2013)
  - #Logic Tx: **1.6× / node** (= 1.57B in 2013)
  - #Cores: **1.26× / node** (= 4 in 2013)
  - #Logic Tx per core: **1.26× / node** (= 0.39B in 2013)
  - # SRAM Tx: **1.6× / node** (= 0.83B in 2013)
- **[NEW]** A-factor
  - Logic (NAND2): **155**
  - SRAM: **60** (bulk and FinFET)
- **[NEW]** Overheads
  - Logic overhead ($O_{eq-logic}$): **1.64 in 2013, scales at 1.26× / 3 years**
  - SRAM overhead (O_{SRAM}): **1.3 from 2013 to 2019, scales at 1.26× / node from 2020**
  - Integration overhead ($O_{integration}$): **1.425** (fixed throughout roadmap)
- **[NEW]** Design-based equivalent scaling for logic
  - 6y-DES: 1.00 in 2013, scales at **0.93× / year** from 2013 to 2019 (to recover one node of Tx scaling at the end of 6 years)**, 0.63** from 2020 onwards
- Power
  - Frequency: **1.04× / year** (same as ITRS 2011) (= 2.0GHz in 2013)
  - Activity factor: **0.95× / year** (same as ITRS 2011) (= 0.07 in 2013)
  - Ratio of low Vt cells (β): 0.1 (same as ITRS 2011)
  - Derated activity factor for non-critical paths (α'): 0.33 (same as ITRS 2011)
  - **[NEW]** Weighted activity factor[4]: **0.32 in 2013** (fixed throughout the roadmap)

[4] *Only 10% of the SOC blocks run at Fmax and the remaining typically run at 1/4th of Fmax. Weighted activity factor is used to obtain final activity factor of the SOC-CP product.*